

# Machine Learning Models implemented into GUI Application for Accurate Prediction of Health Insurance Charges

MScResearchProject  
MScinDataAnalytics

**Bhagyashree M Kenche**  
StudentID:x22228233

SchoolofComputing  
NationalCollegeofIreland

Supervisor: Rejwanul Haque

**National College of Ireland  
Project Submission Sheet  
School of Computing**



<b>Student Name:</b>	Bhagyashree M Kenche
<b>Student ID:</b>	x22228233
<b>Programme:</b>	MSc in Data Analytics
<b>Year:</b>	2024
<b>Module:</b>	MSc Research Project
<b>Supervisor:</b>	Rejwanul Haque
<b>Submission Due Date:</b>	02/09/2024
<b>Project Title:</b>	Machine Learning Models implemented into GUI Application for Accurate Prediction of Health Insurance Charges.
<b>Word Count:</b>	6364
<b>Page Count:</b>	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

<b>Signature:</b>	Bhagyashree M Kenche
<b>Date:</b>	2nd September 2024

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

Attach a completed copy of this sheet to each project (including multiple copies).	<input type="checkbox"/>
<b>Attach a Moodle submission receipt of the online project submission</b> , to each project (including multiple copies).	<input type="checkbox"/>
<b>You must ensure that you retain a HARD COPY of the project</b> , both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

<b>Office Use Only</b>	
Signature:	
Date:	
Penalty Applied (if applicable):	

# Machine Learning Models implemented into GUI Application for Accurate Prediction of Health Insurance Charges.

Bhagyashree M Kenche  
x22228233

## Abstract

Health Insurance plays the most important roles in lives for sustainability. This industry being the global need faces significant challenges as it lags in providing the insurance charges as per the customers need. So, in-order to bridge this gap between raw data and actionable insights, allowing the insurance providers to optimize their operations and pricing at the same time encouraging and empowering policyholders to take better informed decisions about their insurance and healthcare coverage. This project aims at creating a graphical user interface (GUI) Application for predicting the Health Insurance Charges in real time with significant percentage of accuracy. Analyzation of various features such as age, sex, bmi (body mass index), smoker status, number of children, region helps to study and evaluate the performance of different models, including linear regression, random forest, and ensemble models. The outcomes display the effectiveness of these models, precisely in detecting the most influential factors in predicting costs. This work is evaluated using metrics like Mean Absolute Error, Mean Squared Error and R-squared value. Additionally, the research highlights the necessity of data visualization in understanding patterns and distributions within dataset. Moreover, the study thrives to contribute as a basis towards improving real time estimation tools for health insurance costs, leading to spotlight the decision making process and improve health insurance transparency for individuals in Ireland and beyond. Furthermore, this project explores the reliability and potential improvements in the predictive modelling process to better serve the insurance industry.

**Keywords:** Health Insurance, Charges prediction, Machine Learning, Feature Engineering, Model Evaluation, Insurance Industry.

## 1 Introduction

Recently, the increasing complexity of health insurance plans and the rising costs of healthcare services have created a significant challenge for individuals looking to make informed decisions about their healthcare coverage. Ireland private health insurance having long waiting lists makes it crucial to ensure access to preferred healthcare providers, understanding and predicting insurance costs. This project highlights the necessity of accuracy, real-time estimation of health insurance cost by developing a predictive model based-on demographic and health related factors. With Utilization of ML(Machine Learning) techniques, this project goals to empower consumers with the data they need to make informed decisions, thereby enhancing transparency and accessibility in the private health insurance market. The project is

aiming to provide a reliable tool that can simplify the decision-making process and contribute to more equitable and efficient healthcare coverage.

## **1.1 Motivation**

With growing healthcare expenses globally, insurance industry come across rising pressure to estimate costs more accurately to manage risks, better design insurance products, and give competitive pricing. The increasing need précised predictions of healthcare costs, which is critical for both insurance providers and policyholders motivates for this project. As accurate cost predictions empowers patients to make informed decisions on their healthcare options, in turn leading to better financial planning and optimal economic strain. Here, the dataset is used as it includes a wide range of patient data namely, age, sex, bmi (body mass index), smoker status, number of children, region, all of which are critical in delivering the cost. Being comprehensive this data boosts for the development of a robust model that can identify complex relationship between these variables and the resulting insurance costs. Models used in here are Linear Regression, Random Forest, Gradient boosting, and Ridge regression to explore linear and non linear relationships within data. To handle high dimensional data and capture minute patterns through ensemble learning, random Forest and Gradient boosting models is being trained. Cross validation and grid search for hyperparameter tuning, were deployed to make sure model displays reliable and generalized outputs. Additionally, metrics considered in this are MSE (mean squared error), MAE (mean absolute error) and R squared display more insights into the predictive model's performance. By implementing the Machine Learning (ML) models, development of predictive model that leads accurate cost estimates, which can offer more personalized insurance plans and better financial planning for patients.

## **1.2 Research Objectives**

- To develop an efficient predictive model by outstretching the usability of AI in insurance industry.
- Carry out feature importance analysis to identify which of the attributes have the most effect on the prediction of the insurance cost.
- To check and validate the model's performance around various subsets to prove its reliability and applicability it can be.
- Optimization of the model by deploying Hyperparameter tuning with techniques such as Grid Search and cross validation mainly focusing on Random forest and Gradient Boosting models to upgrade its accuracy and efficiency.

## **1.3 Research Question**

- How precisely and reliably can a machine learning model predict health insurance costs based on person's age, sex, bmi (body mass index), smoker status, number of children, region data, and what are the key factors influencing these predictions that can be improved to improve cost estimation practices within the healthcare insurance industry?

## 2 Related Work

Predicting health insurance cost is a demanding crucial task as healthcare expenditure continues to rise globally. Machine Learning models offer significant tools to explore these expenses by analysing big datasets and capture patterns that traditional statistical methods could not. The goal of the projects leans not only to anticipate further cost but also to address strategic decisions, equalize pricing and develop patient care by identifying at-risk individuals. The publications reviewed below gives a broader overview of various machine learning approaches ranging from traditional regression models to advanced ensemble methods and neural networks, underlining the progression of predictive modelling in healthcare insurance industry.

### 2.1 Traditional and Modern Regression Techniques used in predictions in the Recent Publications.

Many of the few papers focused on traditional regression models, such as Linear regression, ridge Regression, Lasso regression and Polynomial Regression. Models like these are well-established in statistical analysis and have been widely applied to healthcare data. Christobel., (2022),[8] compares various regression models, demonstrating that Polynomial regression often outperforms the simpler models like Linear Regression due to its ability to capture non-linear relationships in the data. Nevertheless, though these models are comparatively easy to implement and interpret they have boundaries. Traditional regression models assume linear relationship between the dependent and independent variables, that eventually leads to no proper handling with complex healthcare data. Moreover, the impact of such models can be bordered by the quality of the data, particularly in situations or cases where the data is noisy or contained more outliers.

In the paper ‘Analysis of Cost Prediction in Medical Insurance Using Modern regression Models’ Alzoubi.,(2022 October) [3] explores much of sophisticated regression techniques that bring into notice the limitations of traditional models. These are inclusive of Generalized Linear Models (GLMs), that let for a broader range of data distributions, and models like LASSO (Least Absolute Shrinkage and Selection Operator), which can hold High-dimensional data by selecting only the impactful attributes/features. LASSO regression is specifically of high usage in healthcare datasets, where number of potential predictors could be larger. By applying a fine to the coefficients of less important variables, LASSO reduces the risk of overfitting, eventually turning the model into more generalizable to new data. However, though LASSO and likewise models/techniques provide improvements over traditional regression models, they still require significant tuning and validation to make sure their efficiency and effectiveness.

### 2.2 Ensemble Methods for Enhanced Predictive performances from the Existing Studies.

Emerging as a powerful tool Ensemble methods are improving predictive accuracy by combining the strengths of multiple models. In [12] Pfutzenreuter., (2022) survey highlights Random Forest as an effective tool for holding large datasets and identifying non-linear relationships within the data. The survey looked up to comparing Random Forest with traditional linear regression models that lead to finding out that Random Forest offered superior accuracy due to its capability to handle interactions between variable and at the same time

mitigate overfitting through the averaging of multiple decision trees. Another paper addressed by Patidar.,( 2023, January) also credited the Random Forest on its effectiveness. In this study as well performance of Random Forest against Decision Trees and Linear Regression, proving Random Forest outperformed the other models, specifically when concatenated with hyperparameter tuning. This research paper demonstrated that model's ability to aggregate multiple decision trees giving it to generalize better to unseen data, making it a robust choice for predicting health insurance costs.

Boosting methods, specifically XGBoost and Gradient Boosting Machines (GBMs), have been recognized for their high accuracy in predictive modelling. Building models sequentially with these methods with each model correcting the errors of the previous one, will lead to an accountable reduction in bias and variance. In the publication by Pfutzenreuter.,(2022) also the success of XGBoost's gradient boosting framework is highlighted, where adjustments on models are done iteratively to correct errors from previous iterations, allowing it to fulfil the higher accuracy compared to traditional machine learning models. This method's flexibility and power designs it specifically to be more impactful in cases where prediction precision is crucial. Additionally, stacking as an advanced ensemble technique leverages the accountability of each model, potentially leading to even better performance than boosting or bagging alone. In another paper [18] Shakhovska.,(2022) stacking was compared with other ensemble methods such as bagging and boosting. The survey found that stacking generally provided the highest accuracy because it effectively combined the strengths of various models, identifying a broader range of patterns in data. Usage of meta learner has optimized the combination of predictions from various base models, stacking was able to outperform both bagging and boosting in terms of overall predictive performance.

Although ensemble methods have shown superior performance in health insurance cost prediction, their practical application is not without challenges. Mainly, the issue is computational complexity associated with these models, specifically with boosting and stacking, which require significant processing power and memory. To add, ensemble models specially those including multiple layers like stacking, can be less interpretable than simpler models. This can be drawback in healthcare settings where understanding the decision-making process is critical for acquiring trust from stakeholders and ensuring compliance with regulations. Furthermore, the implementation of these models into real-world systems depends on consideration of factors such as scalability, data privacy and security. In the research made by Albalawi.,(2023) there is a discussion on use of Apache Spark as a tool for handling large datasets in ensemble modelling, showcasing that while computational challenges can be mitigated with appropriate tools, they still require significant infrastructure and expertise.

## **2.3 Advanced Computational Intelligence Techniques used in the Current Literatures.**

The Integration of computational intelligence and advanced machine learning has brought evolution in the field of health insurance cost predictions. The approaches include Support Vector Machines (SVMs), neural networks, XGBoost, and hybrid models which can handle complexities of healthcare data such as non-linearity, high dimensionality, and intricate interactions between variables. This Section, discusses advanced techniques that have been

applied in health insurance cost prediction, addressing on insights from several key publications.

Support Vector Machines (SVMs) proving to be the powerful supervised learning models that can be used for both regression task. They are particularly well-suited for datasets where the relationships between variables is not linear. In [11] ul Hassan., (2021) survey tested SVMs but other advanced models like Random Forest, XGBoost outperformed it with Rsquared score of 87.92% and 88.36%. However this highlighted the challenges in tuning SVMs for large and complex datasets, which could lead to lower performance compared to ensemble methods. Deep Learning models in Neural Networks prominence due to its ability to model highly complex and nonlinear relationships in data as discussed in research work by [6](Goundar., 2020) where both Feed Forward Neural Networks(FNNs) and Recurrent Neural Networks(RNNs) for predicting health insurance claims with metrics of accuracy on Training set :93% and 90.38%, testing set:87.85% and 93.58% making it most suitable for modelling time-dependent data such as patients' medical history or insurance claim record.

XGBoost and Gradient Boosting Machines (GBMs) being a scalable and efficient implementation of gradient boosting has emerged as the one of the most popular machine learning algorithms in predictive modelling. It is particularly known for its high accuracy, speed, and flexibility in handling various types of data including healthcare data. The paper published by Pfutzenreuter., (2022) again highlighted its superior performance in predicting health insurance costs, with R squared of 88%. The ability of the model to handle missing data, applying regularization to prevent overfitting, and mange large datasets makes it an ideal choice in such scenarios. Similarly, in a paper by [11] ul Hassan., (2021) displays 88.36% as R square test and RMSE of 0.34 as the XGBoost performance. This framework upgrades the model by concentrating on the error of previous iterations offering it to reach high predictive accuracy with comparatively low computational cost to other deep learning methods. [9] Patidar., (2023, January.) used Random Forest that proved to be the most accurate model with accuracy of 86.29%, R2 score of 0.86 and RMSE of 4841.88.

## **2.4 Big Data Tools and Practical Implementation Challenges faced in the Existing Research Papers.**

Usage of Machine Learning models in predicting the health insurance costs is a complex task that often requires processing large and diverse datasets. Since the datasets grows in size and complexity, the need for robust big data tools and the practical implementation of such models becomes increasingly important. As in paper produced by [21] Xie., (2015) presents a method for predicting the number of hospitalization days using large scale health insurance data. Although the focus is on stays rather than cost this paper uses regression decision tree algorithm but might limit its potential in accuracy. In contrast, Albalawi..., (2023) addresses the application of Machine Learning algorithm to predict healthcare insurance costs using data from the Kaggle repository. Comparing of different regression models including linear regression and polynomial regression to determine which provides the best fit for the data is taken place. In addition, exploration of Apache Spark for big data analysis for handling large dataset is done offering a reliable solution that can handle large volumes of insurance data efficiency.

[22] Lee.,(2018) studied a vast dataset from the National Health Insurance Service (NHIS) in Korea, covering over 5.7 million individuals. This dataset was used to develop a 10-year stroke prediction model using Cox's proportional hazards regression model. The required analysis and model building were conducted using SAS 9.2 a statistical software package which could handle large datasets and complex models. The integration and management of such a large dataset were major challenges. Ensuring the model's accuracy and validity across a dataset of this size required robust data processing and validation techniques. The study also faced challenges in categorizing the stroke risk for individuals into different risk groups and providing personalized health advice based on these predictions. The model had to balance the complexity of the data with the need for timely and accurate risk predictions. The model achieved AUC values of 0.83 for men and 0.82 for women, indicating high predictive power. However, the model's discrimination capability (as measured by the AUC) required careful calibration, especially given the diverse and extensive dataset. The study noted the limitation of not distinguishing between different types of strokes (ischemic vs. haemorrhagic), which could have provided more granular and accurate predictions but would have required even more complex data and models.

## **2.5 Downfalls Identified in the Existing Related Works.**

As in the existing publications several challenges are reflected in data management, model complexity and practical application. In the paper of [17] Mladenovic., (2020) usage of Adaptive Neuro-Fuzzy Inference System (ANFIS) to predict health insurance cost was carried out. Although ANFIS effectively handles nonlinear relationship in data, the model's complexity and computational intensity offered significant challenges. The study's reliability on the relatively small dataset (1338 records) boundaries its scalability and optimization to larger datasets. Additionally, the difference between model accuracy and computational efficiency is apparent, as to rise in number of input variables improvisation of accuracy but at the cost of higher computational cost. This raises questions on its practical applicability in real-world scenarios where computational resources could be limited. In the survey done by [22] Lee., (2018) the use of huge dataset from National Health Insurance Service (NHIS) in Korea(over 5.7 million individuals) presents both potency and difficulties. The survey's reliance on Cox's proportional hazards regression model offers robust predictive capabilities, as evidenced by the high AUC values achieved. However, the research fails to differentiate between the types of strokes that highlights difficulty of balancing the need for detailed, individualized risk predictions with the practical constraints of processing and interpreting large volumes data. [20] Dutta., (2021April) likewise all other literatures reviewed presents drawback in computational intensity and model optimization. In this study concentrates on optimizing prediction accuracy at the expense of model interpretability which is a very crucial factor in healthcare domain where transparent decision-making processes is essential. This lack of focus on model transparency and interpretability limits the practical utility of the models in real-world healthcare settings, where clear and understandable outcomes are necessary for informed decision-making.

Recently, [5]Orji., (2024) employed advanced ensemble models like XGBoost, Random forest and Gradient Boosting Machine to predict medical insurance cost. This research faced challenges related to the computational resources required for tuning and running these models.



The need for Explainable AI (XAI) methods to interpret the results further complicates the practical implementation, particularly in environments with limited computational power. This reliance on complex models and the associated interpretative tools raises questions about the model's scalability and accessibility in more resource-constrained settings, where simpler models might be preferred despite their lower accuracy.

## 2.6 How does this Project Overcome the Existing Research Challenges?

This project successfully identifies the loopholes in the reviewed literature by leveraging modern ML practices, robust feature engineering and significant model evaluation techniques. Unlike the computational intensity and complexity of the models used previously in literatures like ANFIS, it uses optimized and less complex models like Random Forest, Gradient Boosting and Ridge Regression which eventually sharpens the accuracy and efficiency of the model. The scalability and data management issues faced in large datasets, such as those in the NHIS stroke prediction model, are taken care with preprocessing pipelines automatically making the approach more adaptable to larger datasets. Moreover, the project prioritizes both accuracy and interpretability by combining interpretable models with complex ones and using feature importance metrics to make the predictions understandable, a critical concern in the literature. Finally, the systematic hyperparameter tuning optimize the models for real-time applications, ensuring robustness and practicality across various scenarios, effectively overcoming the limitations of model selection and tuning highlighted in the literature.

## 3 Methodology

Healthcare Insurance industry being the most important role player in Ireland for overall healthcare landscape, providing essential services that complement the public healthcare system. Due to the rising cost and aging population, there is a pressure on both individuals and healthcare systems. Precised prediction models can help in managing this expense more effectively by enabling better financial planning and resource allocation. This project exemplifies how predictive analytics can be applied to real-world healthcare challenges. The flowchart diagram depicts the exact footsteps taken in sculpting this model in producing the relevant customized health insurance cost.

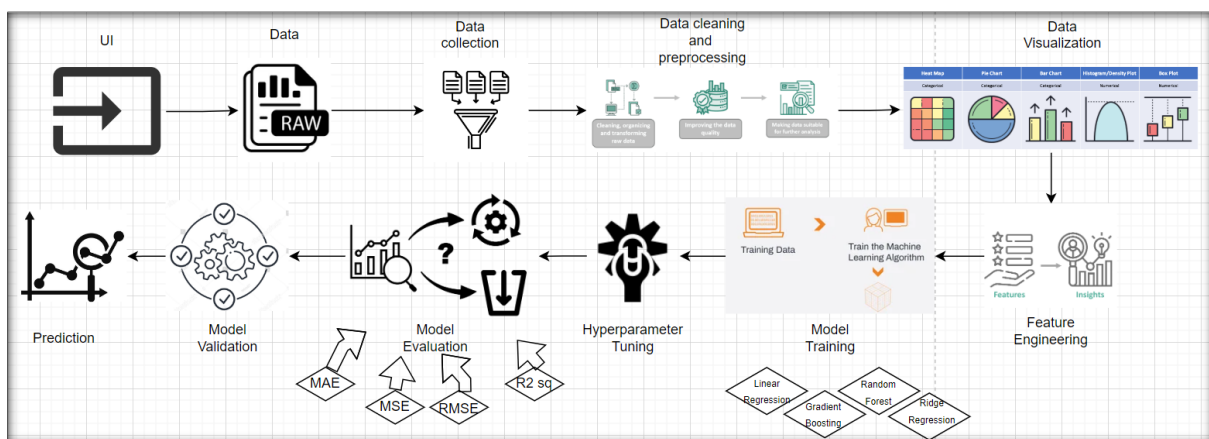


Figure 1: Methodology diagram for prediction of Health insurance cost

### 3.1 Data Description

- Note: -Removed patient demographics in accordance with the GDPR rule.

In this study, the dataset used in here is taken from Kaggle for prediction of health insurance cost that comprises of about 2,772 entries with 7 columns which is used for implementing Machine Learning models —Harish Kumar DataLab. (2023). *Medical Insurance Price Prediction Dataset*. Kaggle.). The dataset consists of 80% for training the model and 20 % for evaluating the model. The size of the dataset is around 112 KB. Key variables consisting in this dataset are age, sex, bmi(body mass index), smoker status, number of children, region.

### 3.2 Data Analysis and Visualization

Data analysis and visualization play vital role in presenting the underlying patterns, distributions, relationships and anomalies data helping in understanding essentials before moving on to Feature Engineering and model building. Examining, cleaning, transforming, and modelling data with the goal of discovering useful information, drawing conclusions, and supporting decision-making are involved in data analysis whereas in Visualization elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Training vs Test Set Distribution in 3D

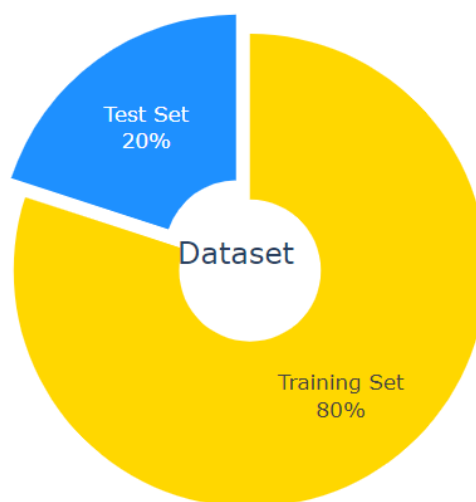


Figure 2: Data Distribution between Train and Test Set

Above pie chart is visualized for the distribution of training and testing data percentage taken from the dataset which is shown in Figure 2. As it shows 80% of the data is used as training data and remaining data is used for testing the model.

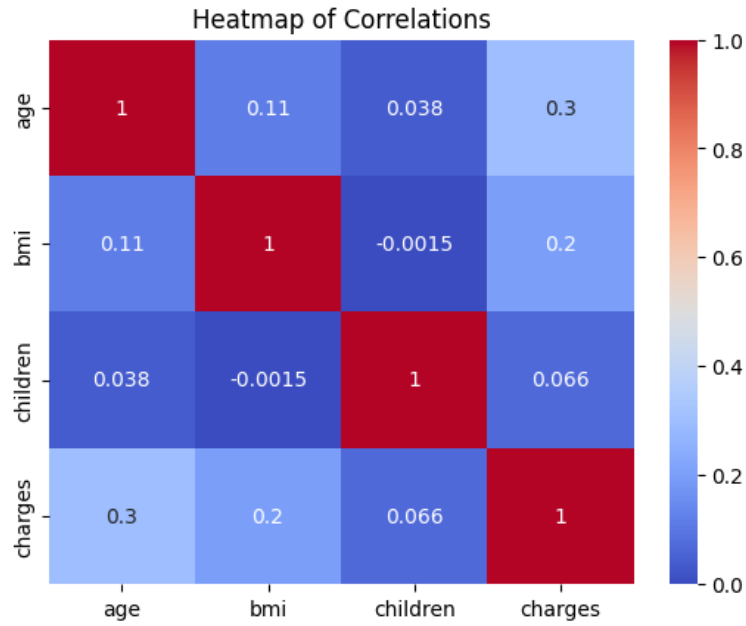


Figure 3: Heatmap of the variables.

In the Figure 3, Heatmap is being visualized for all the variables for it play crucial in providing with a clear, at-a-glance understanding of the relationships within the dataset. This information is very important for making feature selection, deciding model building and interpreting the results. It ensures that the model developed is more accurate and interpretable leading to better insights and predictions.

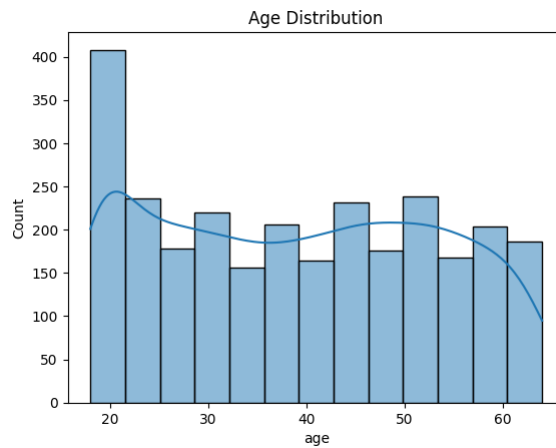


Figure 4: Age Distribution

Here, age distribution is conducted only get an idea about the ranges of age which is visible is from 18 to 64 years, with a mean age of approx. 39.10 years. This distribution suggests a broad range of adult and elderly patients with younger-age group likely being the most prominent. The relatively wide range allows the analysis of how age as a factor influences healthcare costs and outcomes. Similarly, other distributions in Figure 5 are visualized to

understand the large dataset namely- sex distribution ensuring that no bias due to proportionate representation of either gender; BMI distribution; children distribution on how it affect healthcare cost; smoker distribution displays significant impact on overall healthcare insurance costs; region distribution showing wide diversity is critical for improving cost prediction model for setting realistic expectations and budgets.

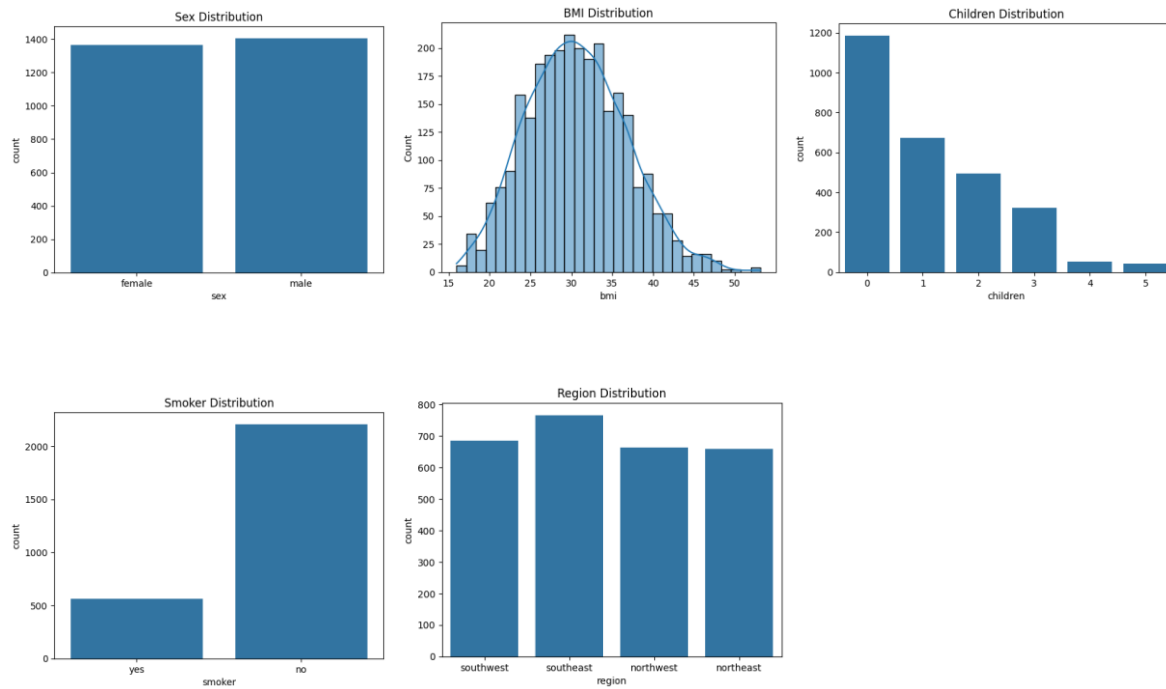
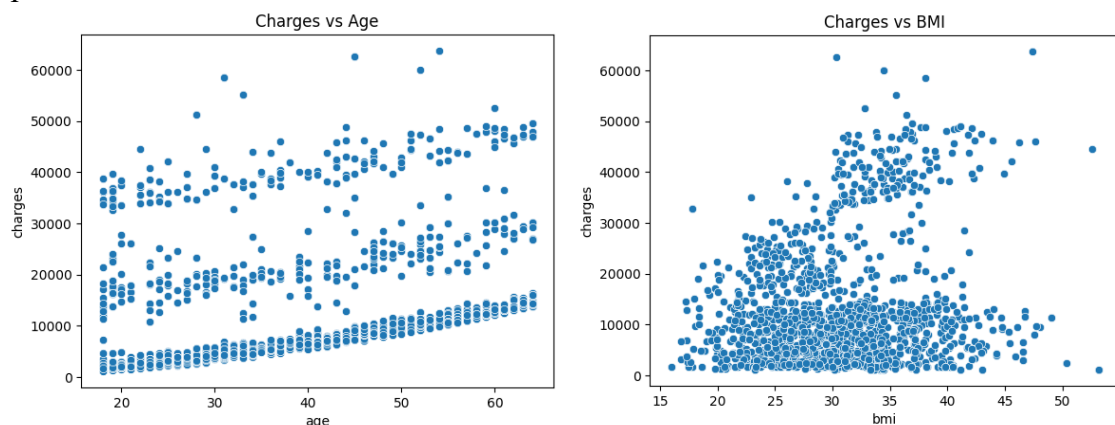


Figure 5: Visualization of distributions of other attributes.

Outliers were also visualized in figure 6 among the attributes mostly situated with charges in the medical insurance data as it can have significant implications on the model training and overall output of the prediction. This will help in increasing the accuracy and reliability of the model. By carefully handling outliers more robust models and good quality of data analysis could be carried out. From the visualization it appears there are several significant outliers which means the dataset consists of few individuals have exceptionally high medical costs compared to the rest of the dataset.



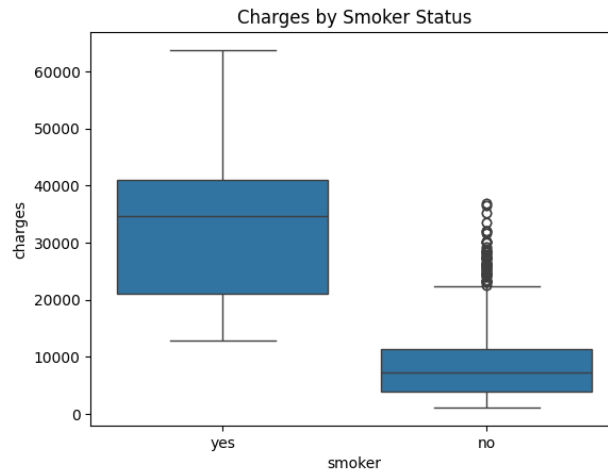


Figure 6: Visualization of outliers in Billing amount.

### 3.3 Feature Engineering

Feature engineering is the crucial process in Machine Learning that involves creating new features or modifying existing ones to improve model performance. In here, categorical variables like sex, smoker and region were transformed using One-Hot Encoding while age, bmi and children were transformed by Standardscaler. Standardizing numerical features and handling missing values using median imputation. Conducting correlation analysis to address multicollinearity, ensuring that the models captured the complex relationships within the dataset, eventually leading to more accurate predictions.

### 3.4 Model Training

Modelling or model training is the basis of the performance and prediction of the results. Under this training the model gets to know how to take in the inputs , work on data, understand various features and its results so that it can perform according when given a new task. In this project, multiple machine learning models, including Linear Regression, random forest, gradient boosting and regression are trained to reach the expected outcome which is predicting health insurance cost. Firstly, the dataset was extracted and split into two sets one 80% for training and rest 20% for testing. After the necessary preprocessing to handle categorical and numerical features established by incorporating steps like One-Hot Encoding, Standardscaler the model was trained and hyperparameter tuning was performed using GridsearchCV to equalize model parameters. Furthermore, the models trained will be discussed step by step in detail in the next Section.

### 3.5 Model Evaluation

Once the model training was carried out successfully next is to check how well is the model performing. This process is called model evaluation which helps in determining whether it accurately captures the underlying patterns of the data or if it is overfitting, underfitting otherwise failing. In this survey, there is a essentiality of evaluation in order to compare the effectiveness of various algorithms like Linear Regression, Random Forest, Gradient Boosting, and Ridge Regression using metrics namely, Mean Absolute Error (MAE), Mean Squared Error

(MSE), Root Mean Squared Error (RMSE), and R-squared(R<sup>2</sup>). With these metrics the magnitude of prediction errors and model's ability to explain the variance in the billing amounts could be figured out. Thus, proper evaluation ensures that the selected model is reliable, robust and suitable for making précised predictions in real world applications.

## 4 Design Specification

This project aims to highlight the necessity for sophisticated models to handle the complex relationships in large healthcare data. below is the in-detail explanation and process followed in employing the Machine Learning Algorithms- Linear regression, Random Forest, Gradient Boosting and Ridge Regression.

### 4.1 Linear Regression

Foundational Algorithm used to predict a continuous target variable by assuming a linear relationship between the input features and the target is Linear Regression. It offers straightforward interpretability, making its easier to understand how much each feature influences the insurance cost. It can be used as a baseline model to compare the performance of more complex algorithm. This algorithm estimates the coefficients(weights) for each input feature such that the sum of squared differences between the observed and predicted values is minimized.

The linear equation can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

where y is the predicted charges x<sub>1</sub>, x<sub>2</sub>,...,x<sub>n</sub> are the input features, and  $\beta_0, \beta_1, \dots, \beta_n$  are the coefficients.

In this project this algorithm provided reasonable predictions but did not properly identify non-linear relationships in the data.

### 4.2 Random Forest

Ensemble Learning model that combines multiple decision trees to improve prediction accuracy and control overfitting is Random Forest. It is working in constructing numerous decision trees during training and averaging their predictions. As this is capable of capturing complex, non-linear relationships between features and the target variable. It surpasses other linear models as it is less sensitive to outliers. It demonstrates the value of different features in predicting the insurance cost.

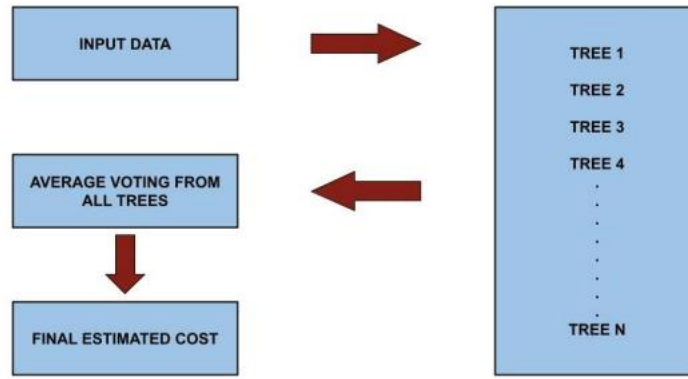


Figure 7: Random Forest Classifier by Patidar, S., & Dudi, S. (2023, January)

Random Forest uses technique called bagging, where multiple subsets of data are created by sampling with replacement. Each subset is used to train a separate decision tree. At each split in the decision trees, a random subset of features is considered for splitting. This introduces additional randomness and helps in reducing the correlation between individual trees. Then predictions from all the trees are averaged to produce the final prediction, which improves robustness and accuracy.

### 4.3 Gradient Boosting

Another Ensemble method which builds models sequentially, where each subsequent model focuses on correcting the errors of the previous ones is Gradient Boosting. This method allows for fine-tuning predictions and improving accuracy. This approach trains model in a sequence then new model focuses on the residual errors ( difference between the actual and predicted values) made by the previous model. The model minimizes a loss function (MSE) using gradient descent, adjusting the predictions iteratively to reduce the errors. Each of the tree in the sequence is weighted based on its performance, and the final prediction is a weighted sum of the individual tree's predictions.

While Gradient Boosting is powerful, it struggled with overfitting in this project, resulting in a performance that was not significantly better than simpler models like Linear Regression.

### 4.4 Ridge Regression

Ridge Regression is a type of linear regression that includes a regularization term to prevent overfitting. It adds a penalty to the magnitude of the coefficients, shrinking them towards zero, which helps in controlling model complexity and improving generalization. This approach modifies the cost function by adding a regularization term:

$$Loss_{Ridge} = \sum (y_i - y^*)^2 + \lambda b^2$$

where  $y^* = a + bX$  is the predicted value.

Figure 8: Ridge regression method by Panda, S., Purkayastha, B., Das, D., Chakraborty, M. and Biswas, S.K., (2022)

The strength of regularization is controlled by the hyperparameter lambda which needs to be tune based on data. However, Ridge Regression did not significantly outperform Linear Regression, suggesting that regularization alone was insufficient to address complexity of the dataset.

## 4.5 Hyperparameter Tuning

In this healthcare insurance billing prediction project, hyperparameter tuning was performed to find the optimal settings for models like Random Forest, gradient Boosting and Ridge Regression. The aim was to enhance the model's performance, improve accuracy, and prevent overfitting. In Random Forest tuning the number of trees(n\_estimators) and the maximum depth of the trees(max\_depth) helps in finding the right balance to increase the predictive performance. Adjusting the regularization parameter in Ridge Regression and controlling the learning rate in Gradient Boosting helped in decreasing overfitting by controlling the complexity of the model.

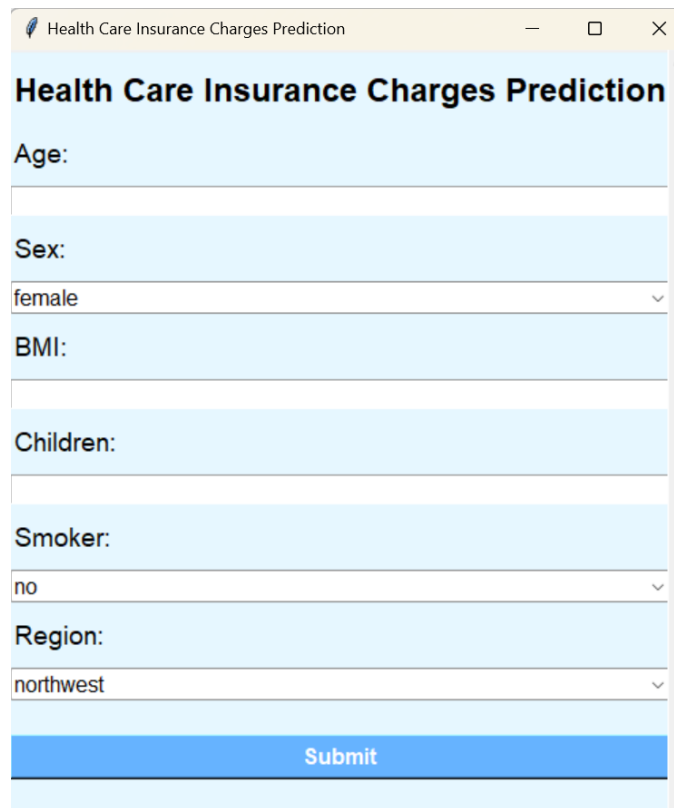
Grid search being an exhaustive search method where the model is trained and evaluated for every combination of hyperparameters specified in a grid was carried out. Cross-validation is used to assess the model's performance on different subsets of the training data. By these two techniques hyperparameter tuning was done. For each model, a range of values for key hyperparameters was defined. GridSearchCV from Scikit-Learn was used to automate the search over this grid of hyperparameters. The cross-validation process helped ensure consistency and generalization of the model's performance.

## 5 Implementation

This paper primarily is created using only one programming language which is Python. After data selection, data cleaning, data preprocess, feature engineering, Model training, Model Evaluation, Model Validation and Prediction, the last stage of implementation involves an interactive graphical user interface (GUI) to take in inputs from the user and return the personalized prediction of health insurance charges in real-time. The application is designed in such a way that it allows users to enter the required data and receive immediate predictions on potential insurance costs using a pre-trained machine learning model. That model is the best model achieved after evaluation and validation process which is Random Forest model. The navigation controls include functionality to navigate between the input form and the result display, enabling users to easily correct inputs and resubmit if necessary. Libraries used to achieve this are tkinter contributed in creating GUI components such as forms, buttons and labels; pandas for handling and converting input data into a format which the model can process and joblib for loading the pre-trained Random Forest model from the particular file. The predictive model was developed using Random Forest algorithm. This model as mentioned

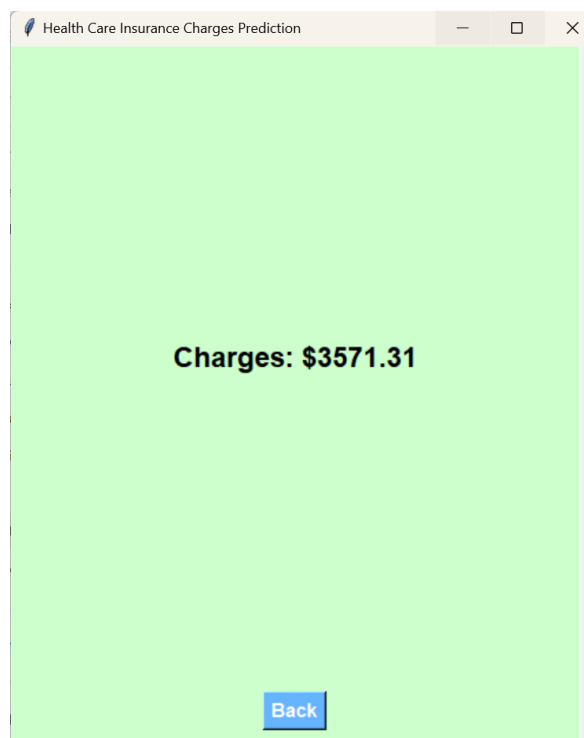


earlier gets trained on the historical data and saves in .pkl file using joblib, which therefore gets loaded and used in GUI application to predict insurance charges.



The screenshot shows a web application window titled "Health Care Insurance Charges Prediction". The interface has a light blue header and a white body. It contains several input fields: "Age:" with a text input, "Sex:" with a dropdown menu showing "female", "BMI:" with a text input, "Children:" with a text input, "Smoker:" with a dropdown menu showing "no", and "Region:" with a dropdown menu showing "northwest". At the bottom, there is a blue "Submit" button.

Figure 9: GUI for inputting data



The screenshot shows the same web application window after a prediction. The input fields are no longer visible. Instead, the main area is a solid light green color with the text "Charges: \$3571.31" in bold black font. At the bottom center, there is a blue "Back" button.

Figure 10: Prediction of Health Insurance Charges

Finally, the desktop application created allows users to type in personal health related information and immediately receive an estimate of their health insurance charges based on a sophisticated machine learning model. Using Python and its libraries enabled the development of a responsive and user-friendly interface while Random Forest model produces robust predictions based on the input data. This solution is enhanced to be intuitive, making users available with actionable insights without requiring any prerequisites on the underlying algorithm or data processing.

## 6 Evaluation

The research where predicting healthcare insurance cost accurately is of foremost importance, several evaluation metrics were deployed to assess the performance of the machine learning models. These metrics include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error(RMSE) and the R squared score. Each of the listed metrics depict unique insights into the model's strengths and weaknesses, helping to identify which models are best suited for deployment and which aspects of the model might require further tuning. Evaluation based-on MSE

Mean Squared Error(MSE) is a metric that measures the average of the squared

differences between predicted and actual values:

$$(MSE) = \frac{\sum_{i=1}^k (E_i - \bar{E}_i)^2}{k}$$

where  $E_i$  are experimental values and  $k$  is total number of data samples.

MSE was used in this reasearch to understand how the model's errors are distributed, especially focusing on larger errors. A high MSE such as the one shown in Linear Regression of 39,963,924 suggests that on average, the squared error is quite large, indicating some significant prediction error. Thus this metric is applied to all the Algoirthms to check models making significant mistakes on certain predictions.

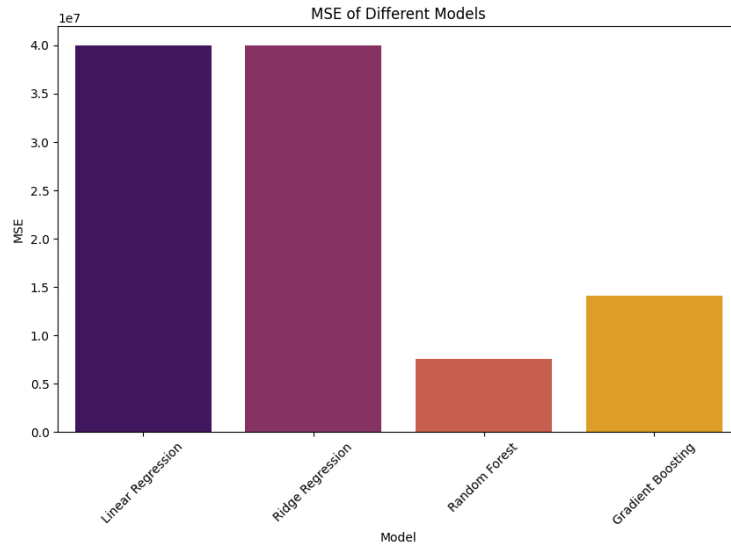


Figure 11: MSE comparison among Machine Learning Models

As in Figure 9 it can be observed that based on MSE metrics Random Forest performed the best with value approximately 75,55,460; Gradient being the second best exhibited value of 14,058,226 & Ridge regression and Linear Regression being almost the same with value of 39,938,747 and 39,963,924 lead at last.

## 6.1 Evaluation based-on MAE

Mean Absolute Error (MAE) is a metric the measures the average magnitude of errors in a set of predictions, without considering their direction (i.e it does not matter whether the prediction was higher or lower than the actual value). It is calculated as the average of the absolute differences between predicted and actual values:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Where  $y_i$  is the actual value and  $\hat{y}_i$  is the predicted value.

In this evaluation the model's prediction was closely matched the actual billing amounts on average. For instance, if a model produced an MAE of 6,194 this means that on average the model's predictions were off by approx. \$6,194 from the actual charges. A lower MAE indicates better model performance as it signifies smaller average errors in the predictions.

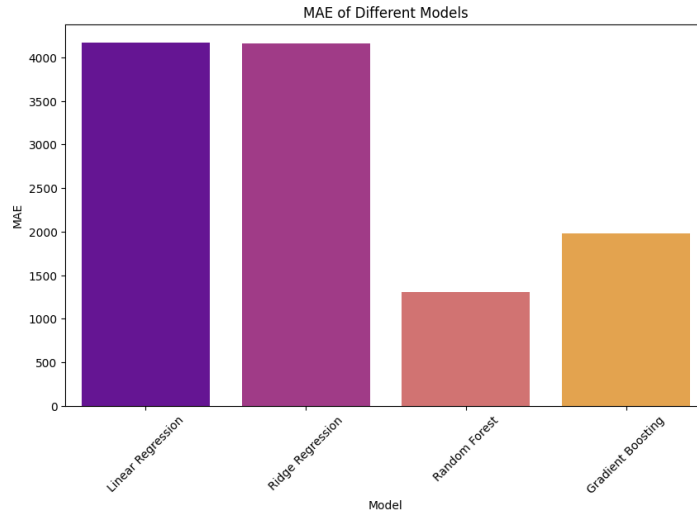


Figure 12: MAE Comparison among Machine Learning Models

In the Figure 12 it is observed Random Forest model achieved the best MAE of approximately 1306, Gradient Boosting showing the MAE of 1977 stands second best, Ridge Regression at third with 4162 not of much difference with the Linear regression at last with highest MAE 4171.

## 6.2 Evaluation based-on RMSE

Root Mean Squared Error (RMSE) is the square root of the mean Squared Error:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

RMSE provides a metric that is in the same units as the target variable, making it easier to interpret compared to MSE. RMSE like MSE is less sensitive to outliers and penalizes large errors more heavily. An RMSE of 4,112 indicates that on average, the prediction errors are around \$4,112. This metric is critical in determining how well the model performs in overall, specifically when considering larger errors. Lower RMSE values across models suggested better performance in predicting billing amounts accurately.

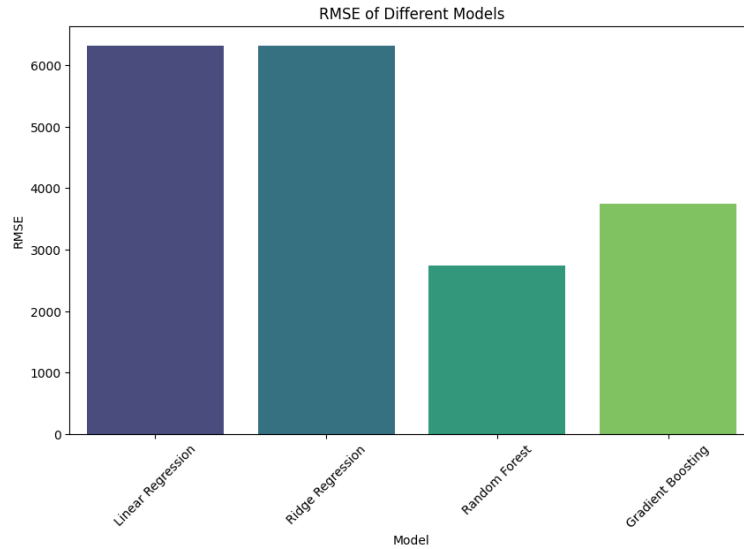


Figure 13: RMSE Comparison among Machine Learning Models

From the Figure 13 it could be derived that lowest RMSE was scored by Random Forest approximately 2,748 proving it to be as the best model, Gradient Boosting of 3,749 RMSE and worst model performance presented by Linear regression and ridge regression showed same RMSE value of 6,321 and 6,319.

### 6.3 Evaluation based-on $R^2$ score

The  $R^2$  score or coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It is calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Where  $\bar{y}$  is the mean of the actual values.

In the project,  $R^2$  score is used to assess how well the model could explain the variance in billing amounts. A non-zero or negative  $R^2$  score, as observed in some models, indicated that the model was performing poorly, essentially no better than random guessing or simply predicting mean billing amount.

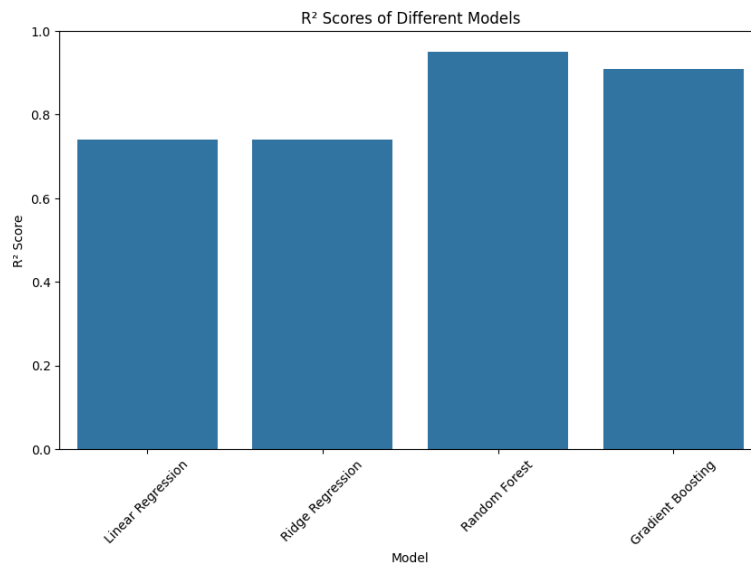


Figure 14: R square Score Comparison among Machine Learning Models

From the above Bar graph in the Figure 14 demonstrate the  $R^2$  score of all the models where Random Forest attains the highest score of 95% along with Gradient Boosting model showing nearby score of 90% leaving the Linear regression and Ridge regression at 73% of  $R^2$  score.

## 6.4 Discussion

In the study, the initial objective was to build machine learning model capable of accurately predicting healthcare insurance charges based on patient demographics including age, sex, bmi (body mass index), smoker status, number of children, region. Four models- Linear regression, random forest, Gradient Boosting and Ridge Regression were evaluated using metrics – Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R square score. It was observed that Random Forest model performed reasonably well in terms of error metrics and score than other models. The Feature importance analysis was carried out in the end to highlight that smoker, bmi and age were the most significant predictors of healthcare insurance amounts, which aligns with the expectations in healthcare settings were older patients and those with longer hospital stays typically incur higher costs. Other important features included Gender, Admission type and Insurance Provider, reflecting their roles in determining the complexity and cost of care.

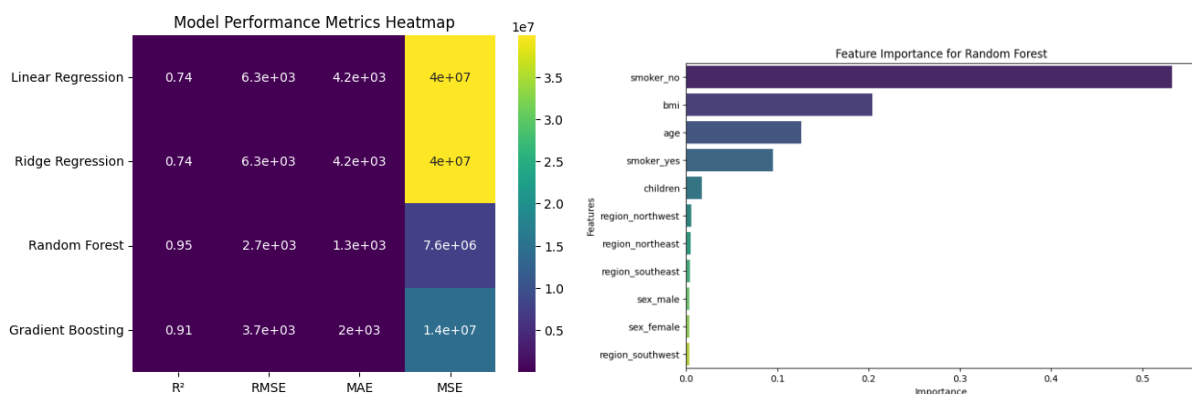


Figure 15: Model Performance Metrics  
Heatmap

Figure 16: Feature Importance for Random  
Forest

## 7 Conclusion and Future Work

The research successfully identified key factors influencing healthcare insurance billing amounts and developed a predictive model with reasonable accuracy. Although Random Forest model performed the best among those tested, the results suggest that there is still significant unexplained variance in the billing data indicating the need for further refinement and additional data. The knowledge acquired from this project gives a foundation for more accurate and comprehensive predictive models in healthcare billing, with the potential to develop financial planning and management in healthcare insurance industry.

Future Work should be focusing on extending the dataset, exploring more complex models, and validating the findings across different contexts to fully realize the potential of predictive modelling in this domain.

## References

- [1] Gregori, M. Petrinco, S. Bo, A. Desideri, F. Merletti and E. Pagano, "Regression models for analyzing costs and their determinants in health care: an introductory review", International Journal for Quality in Health Care, vol. 23, no. 3, pp. 331-341, 2011. **URL:** <https://doi.org/10.1093/intqhc/mzr010>
- [2] Bhatia, K., Gill, S.S., Kamboj, N., Kumar, M. and Bhatia, R.K., 2022, May. Health Insurance Cost Prediction using Machine Learning. In 2022 3rd International Conference for Emerging Technology (INCET) (pp. 1-5). IEEE. **URL:** <https://doi.org/10.1109/INCET54531.2022.9823241>
- [3] Alzoubi, H.M., Sahawneh, N., AlHamad, A.Q., Malik, U., Majid, A. and Atta, A., 2022, October. Analysis Of Cost Prediction In Medical Insurance Using Modern Regression Models. In 2022 International Conference on Cyber Resilience (ICCR) (pp. 1-10). IEEE. **URL:** <https://doi.org/10.1109/ICCR54864.2022.9885895>
- [4] Bhardwaj, N. and Anand, R., 2020. Health insurance amount prediction. Int. J. Eng. Res, 9, pp.1008-1011. **URL:** <https://doi.org/10.17577/IJERTV9IS050778>
- [5] Orji, U. and Ukwandu, E., 2024. Machine learning for an explainable cost prediction of medical insurance. Machine Learning with Applications, 15, p.100516. **URL:** <https://doi.org/10.1016/j.mlwa.2024.100516>

- [6] Goundar, S., Prakash, S., Sadal, P. and Bhardwaj, A., 2020. Health Insurance Claim Prediction Using Artificial Neural Networks. *International Journal of System Dynamics Applications (IJSDA)*, 9(3), pp.40-57. **URL:** <https://doi.org/10.4018/IJSDA.2020070103>
- [7] Chernew, M., Cutler, D.M. and Keenan, P.S., 2005. Increasing health insurance costs and the decline in insurance coverage. *Health services research*, 40(4), pp.1021-1039. **URL:** <https://doi.org/10.1111/j.1475-6773.2005.00412.x>
- [8] Christobel, Y.A. and Subramanian, S., 2022. An Empirical Study Of Machine Learning Regression Models to Predict Health Insurance Cost. *Webology*, 19(2). **URL:** <https://doi.org/10.14704/WEB/V19I2/WEB19242>
- [9] Patidar, S. and Dudi, S., 2023, January. Estimating Medical Insurance Cost using Linear Regression with HyperParameterization, Decision Tree and Random Forest Models. In *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 504-508). IEEE. **URL:** <https://doi.org/10.1109/Confluence54702.2023.10053812>
- [10] Panda, S., Purkayastha, B., Das, D., Chakraborty, M. and Biswas, S.K., 2022, May. Health insurance cost prediction using regression models. In *2022 International conference on machine learning, big data, cloud and parallel computing (COM-IT-CON)* (Vol. 1, pp. 168-173). IEEE. **URL:** <https://doi.org/10.1109/COM-IT-CON53935.2022.9820278>
- [11] Shyamala Devi, M., Swathi, P., Purushotham Reddy, M., Deepak Varma, V., Praveen Kumar Reddy, A., Vivekanandan, S. and Moorthy, P., 2021. Linear and ensembling regression based health cost insurance prediction using machine learning. In *Smart Computing Techniques and Applications: Proceedings of the Fourth International Conference on Smart Computing and Informatics, Volume 2* (pp. 495-503). Springer Singapore. **URL:** [https://doi.org/10.1007/978-981-16-1787-4\\_49](https://doi.org/10.1007/978-981-16-1787-4_49)
- [10] Albalawi, S., Alshahrani, L., Albalawi, N. and Alharbi, R., 2023. Prediction of healthcare insurance costs. *Computers and Informatics*, 3(1), pp.9-18. **URL:** <https://doi.org/10.35834/ci.2023.0004>
- [11] ul Hassan, C.A., Iqbal, J., Hussain, S., AlSalman, H., Mosleh, M.A. and Sajid Ullah, S., 2021. A computational intelligence approach for predicting medical insurance cost. *Mathematical Problems in Engineering*, 2021, pp.1-13. **URL:** <https://doi.org/10.1155/2021/5580401>
- [12] Pfutzenreuter, T. and de Lima, E., 2022. Machine Learning in Healthcare Management for Medical Insurance Cost Prediction. In *OPEN SCIENCE RESEARCH II* (Vol. 2, pp. 1323-1334). Editora Científica Digital. **URL:** <https://doi.org/10.37885/220207863>
- [13] Krishnamurthy, S., Ks, K., Dovgan, E., Luštrek, M., Gradišek Piletič, B., Srinivasan, K., Li, Y.C., Gradišek, A. and Syed-Abdul, S., 2021, May. Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan. In *Healthcare* (Vol. 9, No. 5, p. 546). MDPI. **URL:** <https://doi.org/10.3390/healthcare9050546>
- [14] Akbar, N.A., Sunyoto, A., Arief, M.R. and Caesarendra, W., 2020, November. Improvement of decision tree classifier accuracy for healthcare insurance fraud prediction by using Extreme Gradient Boosting algorithm. In *2020 International conference on informatics*,



*multimedia, cyber and information system (ICIMCIS)* (pp. 110-114). IEEE. URL: <https://doi.org/10.1109/ICIMCIS51614.2020.9354314>

[15] Takeshima, T., Keino, S., Aoki, R., Matsui, T. and Iwasaki, K., 2018. Development of medical cost prediction model based on statistical machine learning using health insurance claims data. *Value in Health*, 21, p.S97. URL: <https://doi.org/10.1016/j.jval.2018.04.672>

[16] Kaushik, K., Bhardwaj, A., Dwivedi, A.D. and Singh, R., 2022. Machine learning-based regression framework to predict health insurance premiums. *International journal of environmental research and public health*, 19(13), p.7898. URL: <https://doi.org/10.3390/ijerph19137898>

[17] Mladenovic, S.S., Milovancevic, M., Mladenovic, I., Petrovic, J., Milovanovic, D., Petković, B., Resic, S. and Barjaktarović, M., 2020. Identification of the important variables for prediction of individual medical costs billed by health insurance. *Technology in Society*, 62, p.101307. URL: <https://doi.org/10.1016/j.techsoc.2020.101307>

[18] Shakhovska, N., Melnykova, N. and Chopiyak, V., 2022. An Ensemble Methods for Medical Insurance Costs Prediction Task. *Computers, Materials & Continua*, 70(2). URL: <https://doi.org/10.32604/cmc.2022.022064>

[19] Billa, M.M. and Nagpal, T., 2024. Medical Insurance Price Prediction Using Machine Learning. *Journal of Electrical Systems*, 20(7s), pp.2270-2279.

[20] Dutta, K., Chandra, S., Gourisaria, M.K. and Harshvardhan, G.M., 2021, April. A data mining based target regression-oriented approach to modelling of health insurance claims. In *2021 5th international conference on computing methodologies and communication (ICCMC)* (pp. 1168-1175). IEEE. URL: <https://doi.org/10.1109/ICCMC51614.2021.9418421>

[21] Xie, Y., Schreier, G., Chang, D.C., Neubauer, S., Liu, Y., Redmond, S.J. and Lovell, N.H., 2015. Predicting days in hospital using health insurance claims. *IEEE journal of biomedical and health informatics*, 19(4), pp.1224-1233. URL: <https://doi.org/10.1109/JBHI.2015.2404824>

[22] Lee, J.W., Lim, H.S., Kim, D.W., Shin, S.A., Kim, J., Yoo, B. and Cho, K.H., 2018. The development and implementation of stroke risk prediction model in National Health Insurance Service's personal health record. *Computer methods and programs in biomedicine*, 153, pp.253-257. URL: <https://doi.org/10.1016/j.cmpb.2017.10.009>

[23] Shreekar, C., Kiran, M., Sumanth, D. and Jeevan, P., 2023. Cost prediction of health insurance. *International Research Journal of Engineering and Technology*, 10(01). URL: <https://doi.org/10.15680/IRJET.V10I1.2023.0101002>